the ECHO DEPository
an NDIIPP-partner project
of the University of Illinois at Urbana-Champaign with OCLC and the Library of Congress

METS, MODS and PREMIS, Oh My!

(and a little MIX and other schema too)

Integrating Digital Library Standards for Interoperability and Preservation

Thomas Habing, thabing@uiuc.edu
Grainger Engineering Library Information Center
University of Illinois at Urbana-Champaign

Presentation Outline

- Brief Background on our Project
- Hub and Spoke METS Profile
  - MODS for descriptive metadata
  - PREMIS for technical and provenance metadata
  - MIX (plus some others) for media-specific technical metadata
- Technical Implementation in Java
- Future Plans

Quick Project Background

NDIIPP ECHODEP[1]

http://ndiipp.uiuc.edu/

- Repository Evaluation
- Tools development
  - Web harvesting and archiving (OCLC's WAW)
  - –** Hub and Spoke interoperability and preservation architecture **
- Preservation Research
  - preserving the authenticity and semantic meaning of digital resources through time.

[1]Exploring Collaborations to Harness Objects in a Digital Environment for Preservation

Hub and Spoke

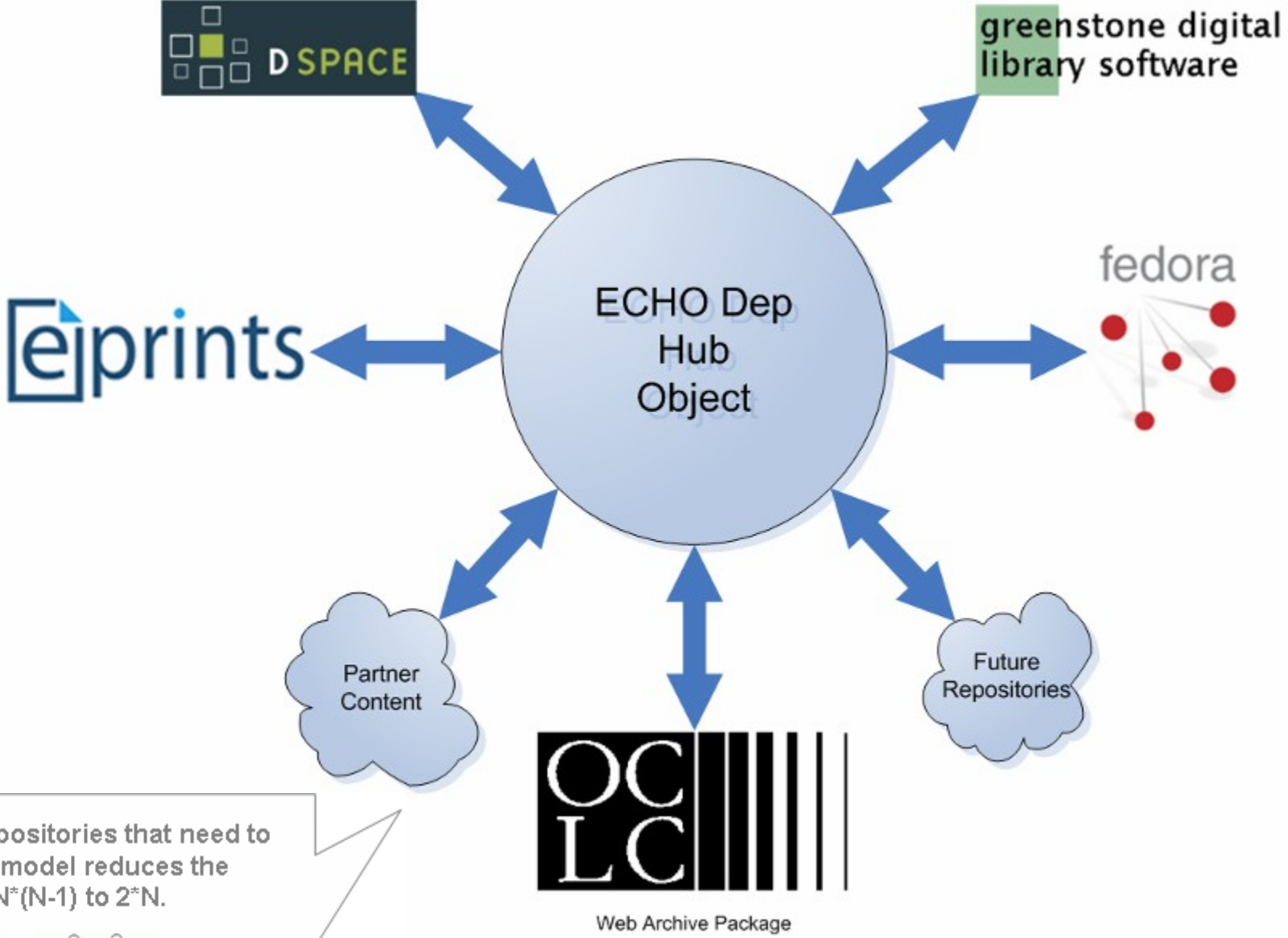
Repository Interoperability Architecture with a forward-looking emphasis on preservation metadata and activities

The Problem

- Plethora of repositories
  - Not just across institutions, but even with a single institution
- Overabundance of data sources
  - Web crawlers like Heritrix or OCLC's WAW, digitization and scanning services, individual authors, batch ingest from legacy systems
- Current integration solutions are local and ad hoc
- Enforcing centralized preservation policy difficult

A Solution

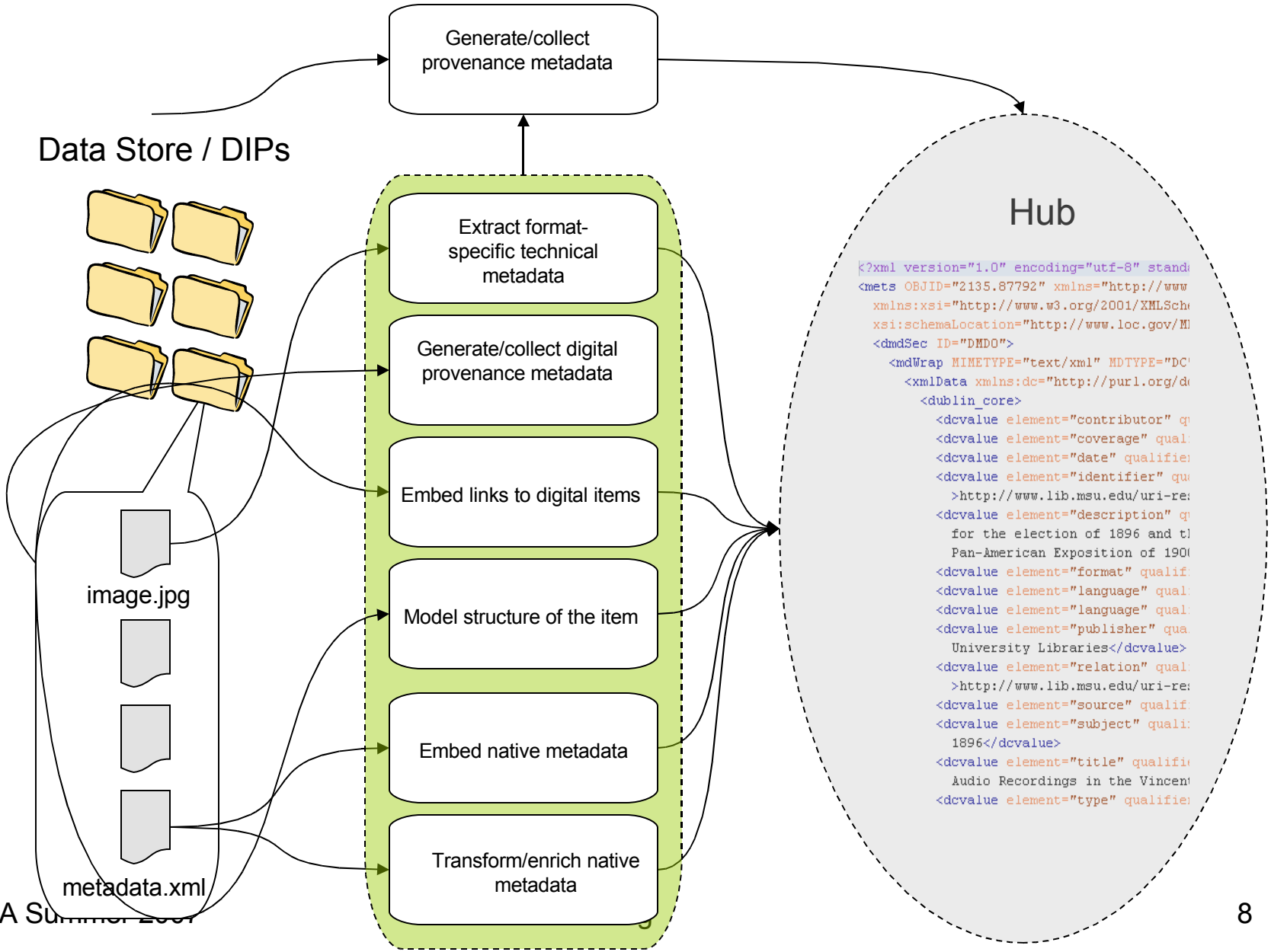
- A common METS-based profile
- A standard programming API
- A series of scripts that use the API and METS profile for creating Information Packages which can be ‘used’ across different repositories



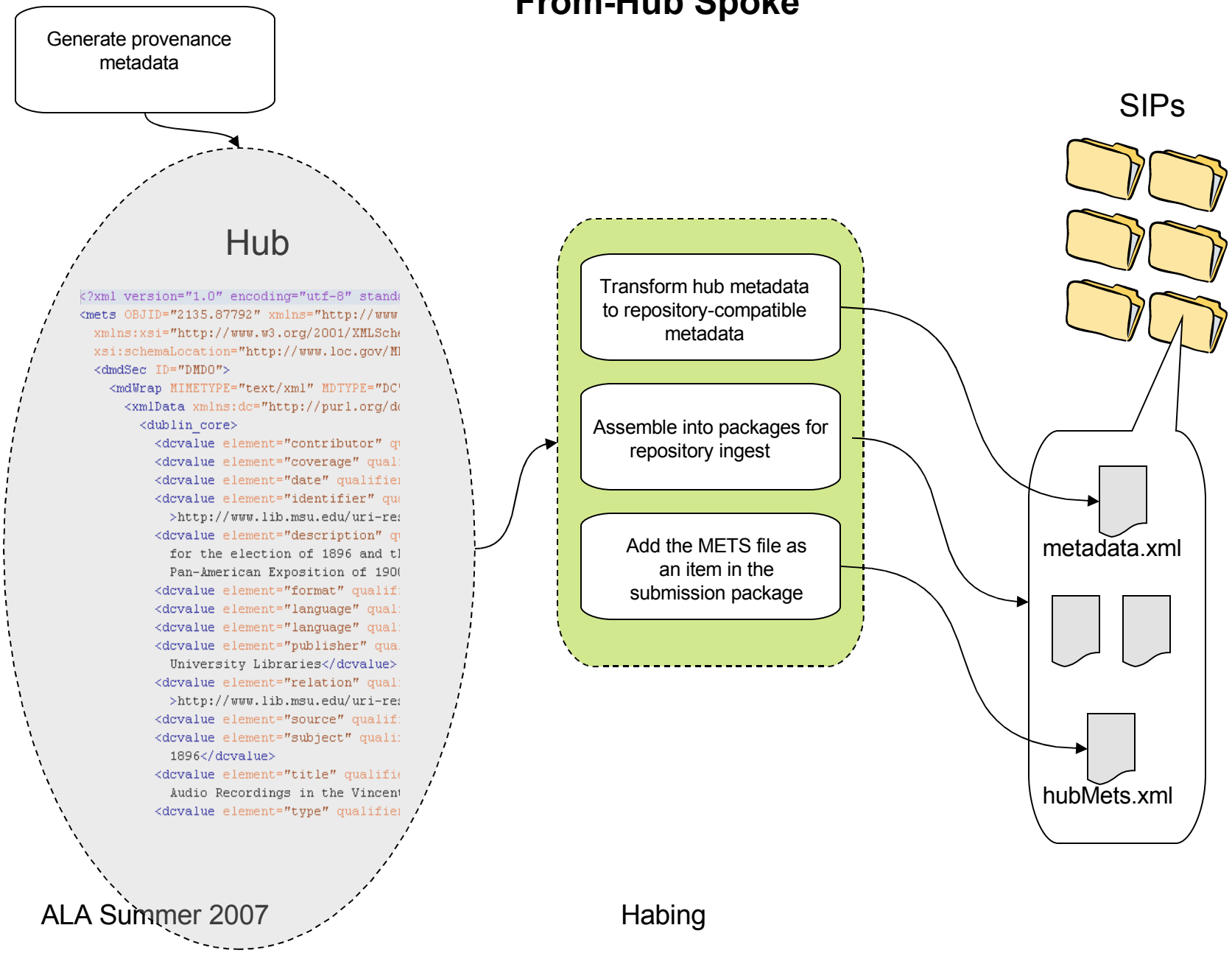
For N different repositories that need to interoperate, this model reduces the complexity from N*(N-1) to 2*N.

This simple idea is the rationale for many different standards that aim to promote interoperability.

To-Hub Spoke


Generate/collect provenance metadata
Data Store / DIPs
Extract format-specific technical metadata
Generate/collect digital provenance metadata
Embed links to digital items
Model structure of the item
Embed native metadata
Transform/enrich native metadata
image.jpg
metadata.xml
Hub
<?xml version="1.0" encoding="utf-8" stand
<mets OBJID="2135.87792" xmlns="http://www
xmlns:xsi="http://www.w3.org/2001/XMLSch
xsi:schemaLocation="http://www.loc.gov/M
<dmdSec ID="DMDO">
<mdWrap MIMETYPE="text/xml" MDTYPE="DC
<xmlData xmlns:dc="http://purl.org/d
<dublin_core>
<dcvalue element="contributor" q
<dcvalue element="coverage" qual
<dcvalue element="date" qualifie
<dcvalue element="identifier" qu
>http://www.lib.msu.edu/uri-re
<dcvalue element="description" q
for the election of 1896 and t
Pan-American Exposition of 190
<dcvalue element="format" qualif
<dcvalue element="language" qual
<dcvalue element="language" qual
<dcvalue element="publisher" qua
University Libraries</dcvalue>
<dcvalue element="relation" qual
>http://www.lib.msu.edu/uri-re
<dcvalue element="source" qualif
<dcvalue element="subject" quali
1896</dcvalue>
<dcvalue element="title" qualifi
Audio Recordings in the Vincen
<dcvalue element="type" qualifie

From-Hub Spoke



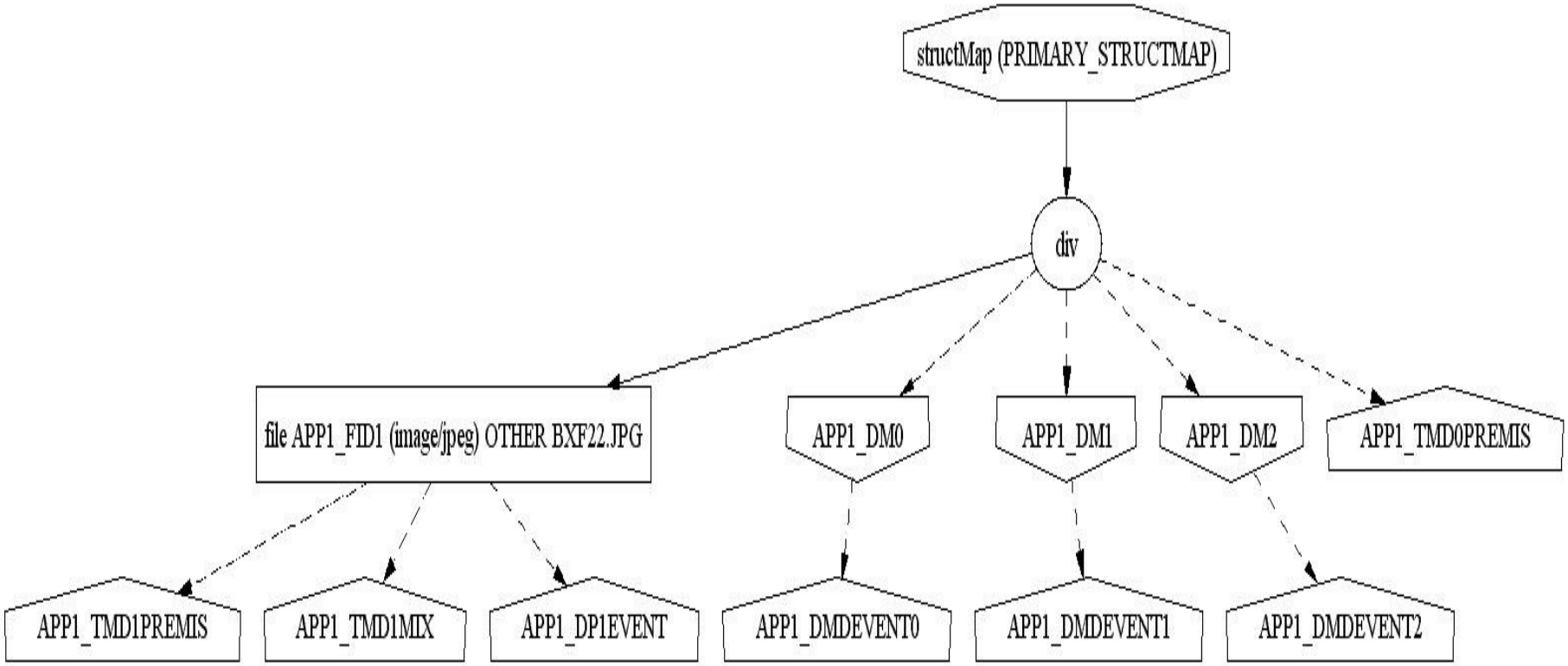
METS Profile

- The METS Profile is the 'Hub'
- Two Registered Profiles
  - http://www.loc.gov/standards/mets/profiles/00000015.xml
    - http://www.loc.gov/standards/mets/profiles/00000016.xml
  - Also http://dli.grainger.uiuc.edu/echodep/METS/
- May be overlaid on top of, or inherited from, other profiles
- Primary Focus of Profiles
  - Digital preservation
  - Repository interoperability
    - minimally at the technical and descriptive metadata level, not at the structural level or file format level
  - Web captures
- Focus on preservation, not access
  - agnostic regarding file formats or structures

METS Profile in More Detail

- Descriptive Metadata
  - Primary DMD is MODS
  - Alternate DMD are encouraged
  - Provenance for DMD is required
- Technical Metadata
  - PREMIS object entities
  - MIX for images
  - Other metadata for other media types
- Digital Provenance
  - PREMIS events and agents

Simple Object Example



http://gita.grainger.uiuc.edu/metsviz/grapher.htm
http://dli.grainger.uiuc.edu/echodep/METS/junit/p1a1.xml

Descriptive Metadata for the Entire Package

- MODS as the primary descriptive metadata
  - The Aquifer MODS profile is used as the minimal requirement (see presentation by Sarah Shreeves)
- Other descriptive metadata schema should be preserved as alternative dmdSec's
- Transformations of descriptive metadata must be recorded in digiprovMD sections using PREMIS event and agent elements
- Individual files may have their own dmdSec's; these are considered outside the scope of our profile. However we encourage the use of relatedItem's in the primary MODS for this purpose.

Technical Metadata for Files

- A techMD section wrapping a PREMIS object element is required for each file or bit-stream
  - Minimal required elements: fixity, size, formatDesignation
  - creatingApplication and software are encouraged especially for MIME types starting with ‘application/…’

Technical Metadata for Files

- Alternative technical metadata schemas for different media types are encouraged:
  - MIX for images
    - http://www.loc.gov/standards/mix/mix.xsd
  - textMD for text
    - http://dlib.nyu.edu/METS/textmd.xsd
  - AUDIOMD for audio
    - http://lcweb2.loc.gov/mets/Schemas/AMD.xsd
  - VIDEOMD for video
    - http://lcweb2.loc.gov/mets/Schemas/VMD.xsd
  - Where possible we are using JHOVE to derive all of these; the profile also allows raw JHOVE output to be used in techMD (http://hul.harvard.edu/jhove/)

Technical Metadata for Representations

- Technical metadata can also be associated with representations
  - There is a special required techMD called the ‘primary representation’ that corresponds to the entire METS file. Used mostly for alternate identifiers for the file, but may also be used to record other technical metadata about the whole METS document
  - Each structural map may also have representation technical metadata.

Digital Provenance

- Recorded for all non-trivial changes to:
  - Descriptive Metadata (must)
    - Creation, Transformation, Modification, Deletion
  - Files and Bitstreams (should)
    - Events from PREMIS data dictionary
  - Structural Maps (may)
    - Creation, Transformation, Modification, Deletion
- PREMIS event and optional associated agents are wrapped in a digiprovMD

Using PREMIS in METS

- All linking via ID & IDREF-type attributes *not* identifier elements
- Embedding
  - Object in techMD
  - Event in digiprovMD
  - Rights in rightsMD
  - Agent in digiprovMD *or* rightsMD
- All Files at a Composition level of 0
  - No packaging, compression, or encryption

Profile for Web Captures

- Inherits almost everything from base profile
- Adds rules for the primary structural map
- Adds rules for referencing ARC files and their constituents from the fileSec
  - ARC is used by Internet Archive, Heritrix web crawler, and OCLC's WAW
  - http://www.archive.org/web/researcher/ArcFileFormat.php

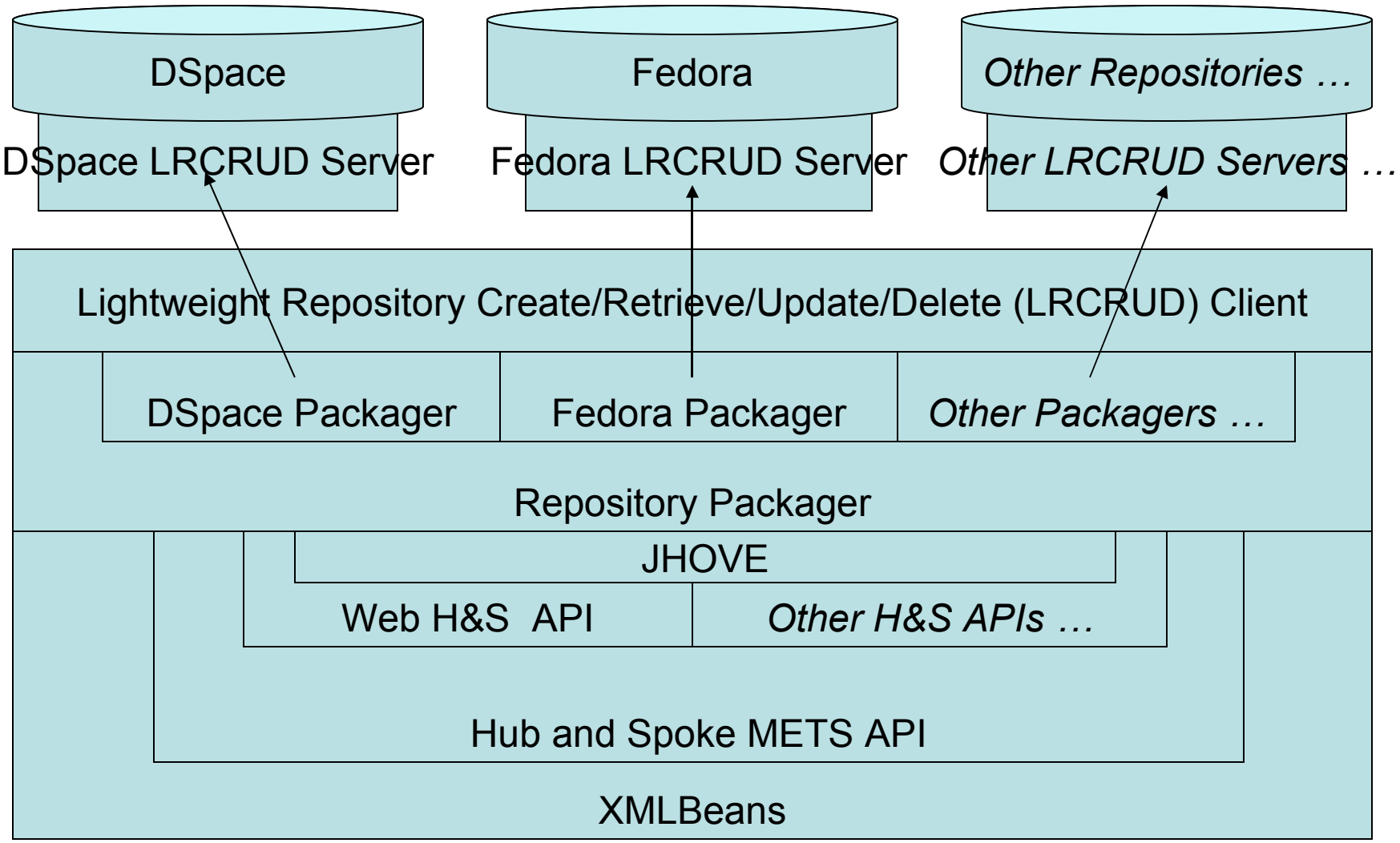
Challenges in Developing the Profile

- How to deal with overlaps between the various schema
  - Properties that occur in multiple places
    - METS attributes, PREMIS elements, MODS elements, MIX elements
  - Differences in how to tie sections together
    - ID and IDREFS or embedded identifiers or nested XML elements
- What METS sections in which to embed the various PREMIS entities

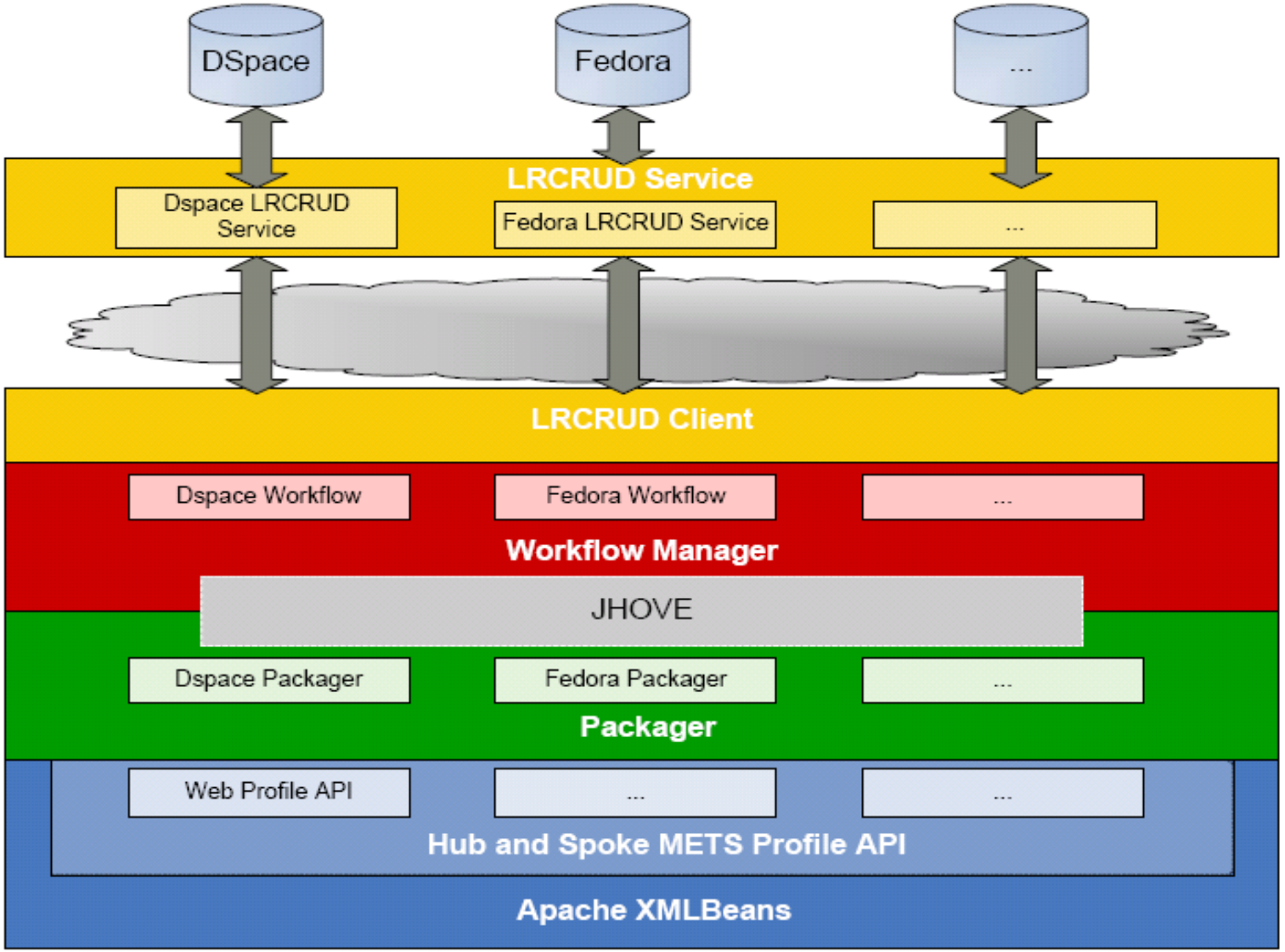
Java Implementation

- Partially complete and in-work
  - Base-level API, plus support for DSpace and to lesser degree Fedora
- Open source
- Javadocs:
  - http://dli.grainger.uiuc.edu/echodep/HnS/JavaDocs/
- Source Code
  - http://sourceforge.net/projects/echodep

Technical Architecture (Java)



Technical Architecture (Java)



Future Plans

- Add support for other repositories such as
  - CONTENTdm
  - EPrints
- Develop additional sub-profiles
- Transformations/Adaptations to/form other METS profiles
- Continue to improve the documentation and program code

Questions?